Title: Characterizing blood modular transcriptional repertoire perturbations in patients with RSV infection: a hands-on workshop using public datasets as a source of training material

Authors: Darawan Rinchai[1], Oceane Konza[2], Signe Hässler[2], Federica Martina[2], Asuncion Mejias[3], Octavio Ramilo[3], Davide Bedognetti[1], Encarnita Mariotti-Ferrandiz[2], David Klatzmann[2] and Damien Chaussabel[1]

Affiliations:

1 Sidra Medicine, Doha, Qatar

2 Inflammation-Immunopathology-Biotherapy Department (i2B), AP-HP, Pitié-Salpêtrière Hospital, Paris, France

3 Division of Infectious Diseases, Nationwide Children's Hospital, Columbus, OH, United States

ABSTRACT

Availability of large collections of public omics data creates a wide range of hands-on training opportunities for biomedical investigators. The organization and outcome of a workshop focusing on implementation of modular repertoire analysis and visualization approaches is described here. This analytic strategy has been developed specifically for mining of blood transcriptome data. Workshop participants conducted modular repertoire analyses of public blood transcriptome datasets generated from patients with acute respiratory syncytial virus infection (RSV). Analyses were conducted using a recently described 382-module repertoire framework. Next, group-level, as well as individual-level modular fingerprint plots were generated. A meta-analysis encompassing 490 subjects was also conducted at the module-level by combining outputs from analyses of six independent datasets. Clustering was used to identify distinct patient subgroups in this consolidated patient cohort.

INTRODUCTION

The availability of vast collections of transcriptome dataset in GEO, the NCBI Gene Expression Omnibus, prompted the development of a training program relying on this resource. The framework for this program was described in a recent review [1]. Briefly, it consists of three training modules: “COD1” focuses on reductionist investigation approaches and relies on transcript profiles of single gene candidates across multiple datasets. In the context of “COD2” trainees curate themed collections of GEO datasets and make them available via a data browsing web application, thus acquiring skills relevant to Omics data management and integration. Finally, “COD3” which is the training module that is directly relevant to the activities described in this paper, consists in developing proficiency in analysis of global scale profiling approaches using public omics data as training material.

Training most specifically relates to the analysis of blood transcriptomics data. Blood transcriptomics consists in measuring abundance of blood leukocyte transcripts on a genome-wide scale [2, 3]. Despite inherent shortcomings, related for instance to the fact that whole blood consists in heterogeneous mix of leukocyte populations, profiling of transcript abundance in whole blood remains one of the most straightforward approach to implement in clinical settings and on a large scale. It has been used to investigate pathogenesis of a wide range of diseases. This includes infectious and autoimmune diseases, but also neurodegenerative or cardiovascular diseases [4, 5]. Responses to interventions, such as vaccines, or antimicrobial therapy have also been assessed via blood transcriptome profiling [6, 7].

The COD3 training modules relies on a specific analytic expertise developed in our laboratory for mining and visualization of omics profiling data. Modular repertoire analysis approach has been described and reviewed previously [8-10]. Briefly the approach is a variation of a co-expression analysis which consist in mining of weighted co-expression networks built on the basis of gene-gene correlation values for a given dataset. For module repertoire construction the network is weighted based on patterns of gene-gene co-expression that are observed across multiple disease or physiological states. For the latest generation of blood modular repertoire used in the context of the analyses presented here a total of 16 different states were used as input, encompassing 985 subjects (Altman et al, manuscript deposition in BioRxiv pending).

The training activities were conducted: 1) on one hand using a private RNAseq dataset, in preparation for subsequent analyses and publication; 2) on the other using public datasets available in the NCBI Gene Expression Omnibus (GEO). The output from the later is being presented in this article. These analyses of public blood transcriptome datasets focused on patients with infections caused by the respiratory syncytial virus (RSV). RSV was selected given its relevance to neonatal health, one of the priority research areas at Sidra medicine.

METHODS

Workshop format

Briefly, a weeklong workshop was organized at Sorbonne University in Paris, France with the participation of two instructors from Sidra Medicine in Doha Qatar. Six

investigators with bioinformatics degrees or with biology degrees and experience in analysis of transcriptome data participated in the training.

On the first day, R scripts developed for modular transcriptome analysis and visualization was shared with the trainees along with a public RSV dataset that served as an illustrative example. A manuscript making the R script in question available along with a detailed description is in preparation, with a draft of the manuscript to be soon available on the BioRxiv pre-print server (Rinchai et al., BioRxiv submission pending).

On the second and third days, modular repertoire analyses were carried on private RNAseq datasets generated in the context of the transimmunom project (described in: [11]). Participants worked individually or in pairs on the same dataset, but using expression matrices that were pre-processed differently, consisting either of raw counts, normalized counts, normalized & batch corrected or normalized, batch and sex-corrected. The result of these and subsequent analyses will be reported in a different manuscript.

On the fourth and fifth days of the workshop public RSV blood transcriptome datasets available in the NCBI gene expression repository GEO were used for hands on activities. The datasets are listed in **Table 1**. Analyses could not conclude by the end of the workshop but reports were submitted afterwards by the participants and are made available here as separate supplementary files. Overall results were compiled and presented in the main text and figures of this article.

Selection of public blood transcriptome datasets

Datasets deposited in the NCBI Gene Expression Omnibus, GEO, were used in the context of this training workshop. Accession IDs along with descriptive information can be found in **Table 1**.

Module repertoire analysis workflow

Analyses were conducted at the “module-level”, with the results being reported for each module as the percentage of constitutive transcripts for which abundance was increased or decreased. When group comparison analyses were performed differences were determined based on a fold change and p-value cutoff. When performing analyses for individual samples the cutoff was based on average of uninfected controls ± 2 SD. This allows adjustment of the fold-change cutoff based on variability observed for each given module across uninfected controls. Output of such analyses are provided for the RSV dataset that was used to illustrate the approach on the first day of the workshop (GSE103842 / Rodriguez-Fernandez et al. [12]).

Results were visualized in a fingerprint format, either as a grid (group level – Figure 1) or as a heatmap (individual level - Figure 5), using again the same illustrative RSV dataset by Rodriguez-Fernandez et al. For each module, % of increased transcripts is represented by a red spot, % decreased transcripts is represented by a blue spot. The largest of the two values is retained for visualization (i.e. there are no purple spots). In the grid format, on Figure 1, the position of each module is fixed. A row of modules corresponds to a “module aggregate” which is a set of modules following a

similar pattern of activity across the 16 input datasets corresponding to the different disease or physiological states. A legend is provided to indicate functional interpretations associated with some of the modules. In the heatmap format, Figure 5, each row corresponds to a module and each column to a sample. Columns and rows are arranged based on similarity in patterns of module activity. Filters can be applied, for instance to remove modules that show only low levels of activity across the samples. Another common filter consists in retaining only the modules associated with functional annotations.

RESULTS

Characterization of the RSV blood transcriptome fingerprint

The fingerprint generated on day one of the workshop for the illustrative dataset from Rodriguez-Fernandez et al (GSE103842; [12]) is shown in **Figure 1**. Functionally, the fingerprint of acute RSV infection is characterized by a marked interferon signature, as illustrated by uniform over-expression of modules associated with aggregate A28 (Figure 1, turquoise color). As is often the case it is accompanied by decrease in abundance of modules associated with protein synthesis (aggregate A1, with blue spots indicating decrease in abundance of constitutive transcripts for modules at positions 1, 5, 23 and 25, in dark purple). It is also accompanied by an increase of modules which are part of aggregate A26 and A27 which are functionally associated with cell-cycle. All these changes have been described previously in systemic Lupus erythematosus, another disease dominated by a pronounced interferon response [9, 13], most recently in the context of the paper describing the third generation modular framework used

during this workshop (see also **Figure 2**, left panel & GSE100150: Altman et al, 2019, pending deposition in BioRxiv). However, a peculiarity of the RSV signature lies with the concomitant presence of signatures associated with immunosuppression (A37 and A38). This is illustrated in **Figure 2**, which shows the fingerprint of RSV for a blood transcriptome dataset from Altman et al., alongside with the fingerprint of SLE and influenza, both dominated by interferon response signatures and the fingerprint of stage IV melanoma and liver transplant recipients under maintenance immunosuppression therapy, both dominated by immunosuppression signatures.

Module-level analyses of six independent RSV blood transcriptome datasets

Trainees were assigned one of the six different public RSV transcriptome datasets listed in **Table 1**. Module fingerprint maps generated for each dataset are shown **Figure 3.** Overall the results confirm the characteristics of the RSV blood transcriptional signature described earlier, with concomitant interferon (row A28) and immunosuppressive signatures (rows A36 to A38) observed in all six datasets. Notably, both of the RSV signatures shown in Figure 1 (Rodriguez-Fernandez et al., GSE103842) and Figure 2 (Altman et al., GSE100150) comprise both interferon and immunosuppressive signatures as well.

Module aggregate-level comparisons across datasets

Next, module aggregate-level profiles were derived from each of the 6 datasets by averaging % values of all modules constituting each aggregate. These profiles were then combined in a single heatmap where module aggregate were arranged as rows

and datasets as columns (**Figure 4**). Datasets were then grouped according to patterns of module aggregate activity. This permitted a more direct comparison of characteristics of the six independent RSV datasets. While the heatmap and sample grouping confirmed increased in abundance for the “interferon” and “immunosuppression” module aggregates across all six datasets, differences were also observed that drove a dichotomy between datasets GSE73072, GSE77087, GSE103842 (“Cluster 1”) on one hand, and GSE80179, GSE42026 and GSE38900 (“Cluster 2”) on the other. Modules aggregates “driving” this separation were chiefly A15, A29 and A30 (collectively with B-cell, cell cycle annotations) which showed marked decreases only in “Cluster 2” datasets and A34 and A36 (collectively with Erythrocyte and platelet annotations) which showed increases again only in “Cluster 2” datasets. A34 showed a modest increase in Melanoma and conversely a decrease in liver transplant recipients and may thus not be obviously associated with immunosuppression (**Figure 2**). A36 on the other hand was increased in both melanoma and transplant, and not in influenza or SLE, which may therefore indicate a possible association with immunosuppression. Some of the factors that could account for differences between the two dataset “clusters” were examined but none could be conclusively associated with one or the other (Table 1: including collection tubes, demographics, array platform). Within each cluster further separation could be observed, most notably for GSE73072 from Liu et al [14]. But differences remained overall relatively minor especially when considering the fact that this dataset was generated for adult subjects exposed experimentally to RSV and using Affymetrix gene chips, which is in contrast with the other five datasets that consisted of profiles of

pediatric subjects with community acquired RSV infection managed in an acute clinical care setting, with transcript abundance data generated on Illumina beadchips.

Modular fingerprinting and classification of individual study subjects.

As part of the training workflow, participants also generated fingerprints of individual subjects within a given dataset which were then displayed on a heatmap where modules and subjects were grouped via hierarchical clustering. An example is given **Figure 5** for the illustrative dataset from Rodriguez-Fernandez et al. (GSE103842) that was used on the first day of the workshop. This representation is useful in that it permits to explore the heterogeneity of the RSV signature across individual subjects. In the case of GSE103842 the subjects were distributed across 5 major groups according to patterns of changes in transcript abundance. The two clusters accounting for most of the RSV subjects were HC4 and HC5. With HC5 subjects characterized by a robust interferon signature accompanied by inflammation and immunosuppression signatures. For HC4 the pattern of module activity is largely similar with the exception of a muted inflammatory signature. HC3 comprised of most of the healthy controls. HC1 and HC2 were relatively minor groups in comparison with only 6 subjects each. HC2 most closely resembled HC4, with low inflammation but a more marked activation of a set of modules associated with cell cycle and relatively muted interferon signature. HC1 on the other hand was uniquely characterized by the presence of a modular signature associated with erythrocytes.

Classification of individual subjects based on differences in abundance patterns of 6 interferon modules.

Subjects from the same illustrative dataset from Rodriguez-Fernandez et al. were used again, while focusing this time on the 6 interferon modules constituting the A28 aggregate. Consistently with prior work which stratified patients with SLE according to distinct interferon module signatures [15], RSV patients could be classified on the basis of the number of interferon modules showing significant increase in abundance of their constitutive transcripts (**Figure 6, top panel**). Subjects belonging to this dataset were partitioned again in 5 distinct cluster corresponding to 5 different branches of the hierarchical clustering tree. The intensity of the interferon signature is indicated as a function of the number of interferon modules showing increase in abundance: In a first group, which encompassed most of the control subjects, this number ranged between 0 and 1. In a second group the range increased with subjects displaying activation of one or two modules. The range further increased in other groups reaching up to 6 modules, including the two modules for which threshold of “activation” appeared to be the highest (M13.17 and M15.64). Overall clinical significance in terms for instance of severity or disease outcome would remain to be further explored. In this particular dataset RSV genotype could for instance be used as information for grouping of patients and examine its association with interferon response (**Figure 6, bottom panel**).

Exploring feasibility of integrative analyses at the module-level.

Finally, output of all module-level analyses conducted for individual subjects across all 6 datasets was combined to evaluate the pertinence of carrying out integrative analyses

encompassing multiple public transcriptome datasets. Percent changes of transcripts constitutive of each of the 382 modules from our third generation blood modular repertoires were compiled across a total of 490 subjects. PCA plots were next generated with color coding of samples according to study group (case vs controls) and dataset membership (GEO GSE identifiers). The results are shown on **Figure 7**. A good separation was observed between RSV cases and controls, as illustrated in the 3D PCA plot (top) and 2 PCA plots (middle). Extensive overlap was on the other hand observed when samples were color-coded according to membership to the each of the six studies (bottom panel). Similar observations could be made when samples were instead plotted via tSNE (t-Distributed Stochastic Neighbor Embedding, [16] – **Figure 8, top and middle panels**). In addition, the 490 samples were partitioned via k-means clustering as a means to identify distinct endotypes of RSV signatures among this large combined sample set (**Figure 8, lower panel)**. Cluster 8 (grey) on the bottom panel appeared to most closely "align" with a core set of healthy controls (top panel) in which all six datasets appear to contribute (middle panel). This cluster can then be reasonably considered to delineate a consensus uninfected baseline. Cluster 7 (green) and cluster 7 (beige) were relatively small in size and appear to encompassed outlying cases and controls. This leaves 5 distinct RSV endotypes (yellow, pink, red, blue, teal) of roughly equal size to further consider from both biological (qualitative and quantitative differences in modular patterns of transcript abundance) and clinical standpoints (e.g. severity).

DISCUSSION

RSV remains a major cause of infant mortality worldwide and has been associated with development of adverse health outcomes later in life, including asthma [17]. To date, efforts towards the development of a vaccine has not proven fruitful [18]. Indeed, this may be in part attributed to immune evasion mechanisms established by the virus that prove able of thwarting the development of long-term immunity against the virus [19]. This may be best illustrated by the fact that, unlike what is observed in the case of influenza virus, it is common for individuals infected by RSV to be subsequently reinfected by the same strain of virus. Of potential interest the observations reported here suggest that subversion of the immune response by the virus may be reflected by signatures bearing a striking resemblance with immunosuppressive states observed in patients with stage IV transplantation or liver transplant recipient receiving immunosuppressive drug regimen. Another lesson that may be derived from these analyses is that while some degree of generalization is helpful when interpreting overall responses, it is also important to recognize that heterogeneity is present when profiling responses at the level of individual patients. This may reflect different degrees of severity or stages in development of the disease. The presence or absence of an inflammation signatures is for instance one of several distinctive features that would need to be further investigated.

Another relevant observation that can be made through the work presented herein is that working with data reduced to the level of genesets may provide an adequate approach to integrating the transcriptome profiles of multiple studies in a single analysis.

An obvious advantage is the ability to work with much larger sets of patients. However, interpretability may remain limited due to the fact that the patient information may not be collected in the same manner and could prove difficult to harmonize “after the fact”.

Finally, organization of the workshop upon which the analyses presented in this article are based provided an opportunity to put to the test a given format for the third training module of the three part “COD” training program currently being developed at Sidra (COD3 = global scale analyses and visualization approaches using public omics data as source of training material – [1]). About only have the time during this weeklong workshop was dedicated to re-analysis of public RSV data and it did not prove sufficient for trainees to complete the analyses during the time imparted. Follow up via email and completion of the assignments as homework proved nevertheless an opportunity for participants to practice and maybe better “integrate” afterwards the know how necessary to carry out such module repertoire analyses.

REFERENCES

1. Chaussabel, D. and D. Rinchai, *Using 'collective omics data' for biomedical research training.* Immunology, 2018. **155**(1): p. 18-23.
2. Banchereau, R., et al., *Understanding Human Autoimmunity and Autoinflammation Through Transcriptomics.* Annu Rev Immunol, 2017. **35**: p. 337-370.
3. Chaussabel, D., *Assessment of immune status using blood transcriptomics and potential implications for global health.* Semin Immunol, 2015. **27**(1): p. 58-66.
4. Devaux, Y., *Transcriptome of blood cells as a reservoir of cardiovascular biomarkers.* Biochim Biophys Acta Mol Cell Res, 2017. **1864**(1): p. 209-216.
5. Song, F., et al., *Perturbation of the transcriptome: implications of the innate immune system in Alzheimer's disease.* Curr Opin Pharmacol, 2016. **26**: p. 47-53.
6. Goletti, D., et al., *Update on tuberculosis biomarkers: From correlates of risk, to correlates of active disease and of cure from disease.* Respirology, 2018. **23**(5): p. 455-466.

7. Nakaya, H.I., S. Li, and B. Pulendran, *Systems vaccinology: learning to compute the behavior of vaccine induced immunity.* Wiley Interdiscip Rev Syst Biol Med, 2012. **4**(2): p. 193-205.
8. Obermoser, G., et al., *Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines.* Immunity, 2013. **38**(4): p. 831-44.
9. Chaussabel, D., et al., *A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus.* Immunity, 2008. **29**(1): p. 150-64.
10. Chaussabel, D. and N. Baldwin, *Democratizing systems immunology with modular transcriptional repertoire analyses.* Nat Rev Immunol, 2014. **14**(4): p. 271-80.
11. Lorenzon, R., et al., *Clinical and multi-omics cross-phenotyping of patients with autoimmune and autoinflammatory diseases: the observational TRANSIMMUNOM protocol.* BMJ Open, 2018. **8**(8): p. e021037.
12. Rodriguez-Fernandez, R., et al., *Respiratory Syncytial Virus Genotypes, Host Immune Profiles, and Disease Severity in Young Children Hospitalized With Bronchiolitis.* J Infect Dis, 2017. **217**(1): p. 24-34.
13. Bennett, L., et al., *Interferon and granulopoiesis signatures in systemic lupus erythematosus blood.* J Exp Med, 2003. **197**(6): p. 711-23.
14. Liu, T.Y., et al., *An individualized predictor of health and disease using paired reference and target samples.* BMC Bioinformatics, 2016. **17**: p. 47.
15. Chiche, L., et al., *Modular transcriptional repertoire analyses of adults with systemic lupus erythematosus reveal distinct type I and type II interferon signatures.* Arthritis Rheumatol, 2014. **66**(6): p. 1583-95.
16. van der Maaten, L.J.P.H., G.E., *Visualizing Data Using t-SNE.* Journal of Machine Learning Research. , 2008. **9**: p. 2579–2605.
17. Stein, R.T., et al., *Respiratory syncytial virus hospitalization and mortality: Systematic review and meta-analysis.* Pediatr Pulmonol, 2017. **52**(4): p. 556-569.
18. Gerretsen, H.E. and C.J. Sande, *Development of respiratory syncytial virus (RSV) vaccines for infants.* J Infect, 2017. **74 Suppl 1**: p. S143-S146.
19. Ascough, S., S. Paterson, and C. Chiu, *Induction and Subversion of Human Protective Immunity: Contrasting Influenza and Respiratory Syncytial Virus.* Front Immunol, 2018. **9**: p. 323.

ACKNOWLEDGEMENTS

Organization of the “COD3” workshop was supported by Inflammation-Immunopathology-Biotherapy Department (i2B), AP-HP, Pitié-Salpêtrière Hospital with the help of Caroline Aheng and Sophie Miller. Sidra Medicine is a member of the Qatar

Foundation for Education, Science and Community Development. The development of the COD training program is supported in part via a grant from the Qatar National Research Fund: QNRF NPRP10-0205-170348

CONTRIBUTIONS

Conceptualization: DR, DC, DB, EM, DK. Data curation and validation: DR, OK, SA, FM Visualization: DR, DC, OK, SA, FM. Analysis and interpretation: DR, OK, SH, FM, AM, OR, EM, DK, DC. Writing of the first draft: DR, DC. Funding acquisition: DK, DC. Methodology development: DR, DC. Writing – review & editing: DR, OK, SH, FM, AM, OR, DB, EM, DK, DC. The contributor's roles listed above follow the Contributor Roles Taxonomy (CRediT) managed by The Consortia Advancing Standards in Research Administration Information (CASRAI) (https://casrai.org/credit/).

FIGURE LEGENDS

Figure 1: Fingerprint grid map of modular repertoire changes in patients with RSV infection vs uninfected controls. Modules are assigned fixed position on a grid, with each row corresponding to a "module aggregate" constituted of modules following similar patterns of changes in transcript abundance. The number of constitutive modules for each aggregate varies between 2 (A16) and 42 (A2). Percentage of constitutive transcripts for a given module showing increase in abundance in RSV over controls is indicated by a red spot. Conversely, percentage of constitutive transcripts showing a decrease in abundance for a given module is indicated by a blue spot. The

color key at the bottom indicate functions which have been associated with some of the modules on the grid.

Figure 2: Fingerprint grid maps of modular repertoire changes in patients with RSV infection, Systemic Lupus Erythematosus, Influenza infection, stage IV melanoma or liver transplants. The visualization scheme employed is similar to the one described for the fingerprint grid map in Figure 1. Since the position of the modules on the grid is fixed the color key from Figure 1 can also be used for functional interpretation. The dataset on which the fingerprint maps are based is available publicly under GEO accession ID GSE100150. Plus signs or NC indicate patterns of activation for module aggregate A28 (interferon modules) and module aggregates A36 through A38 which appear to be associated with tolerogenic or immunosuppressive states.

Figure 3: Fingerprint grid maps of modular repertoire changes in six independent RSV blood transcriptome datasets. The visualization scheme employed is similar to the one described for the fingerprint grid map in Figure 1. Since the position of the modules on the grid is fixed the color key from Figure 1 can also be used for functional interpretation. Maps shown in these figure were generated from multiple independent RSV blood transcriptome datasets which have been deposited in the NCBI's Gene Expression Omnibus.

Figure 4: Fingerprint heat map comparing module aggregate-level patterns of activity across 6 RSV datasets. Module aggregate-level data on this heat map are

arranged in rows and datasets in columns. Datasets are grouped according to similarities in patterns of activity across module aggregates via hierarchical clustering.

Figure 5: Fingerprint heatmap of modular repertoire changes in individual patients with RSV infection and uninfected controls. The dataset employed for this analysis was contributed to GEO by Rodriguez-Fernandez et al. and was attributed the GEO ID GSE103842. Modules on this heat map are arranged in rows and subjects in columns. Ordering is based on similarities in patterns of module activity across rows and columns. Red spots represent the proportion of constitutive transcripts for a given module that are increased. Blue spots represent the proportion of constitutive transcripts that are decreased. A color key is used to indicate study groups (control or RSV) as well as the viral genotype. Blue arrows on the right hand side highlight modules associated with interferon responses belonging to aggregate A28. Orange arrows highlight modules associated with immunosuppression belonging to aggregates A37 or A38.

Figure 6: Fingerprint heatmap displaying patterns of activity of 6 interferon modules across individual study subjects. The heat map on the upper panel displays patterns of activity of 6 interferon modules constitutive of aggregate A28. Subjects are grouped according to similarities in activity for these 6 modules. Groups shown on the heat map indicate the range in number of modules activated for any of the individual in each given group. The color key provides sample information which includes grouping (RSV case or control) and RSV genotype (GA2, GA5, BA). The heat

map in the lower panel displays patterns of activity of the same 6 interferon modules with in this case subjects consolidated into 5 groups, according to RSV genotype. Module order in both heat maps is the same with module identified being provided in the lower panel.

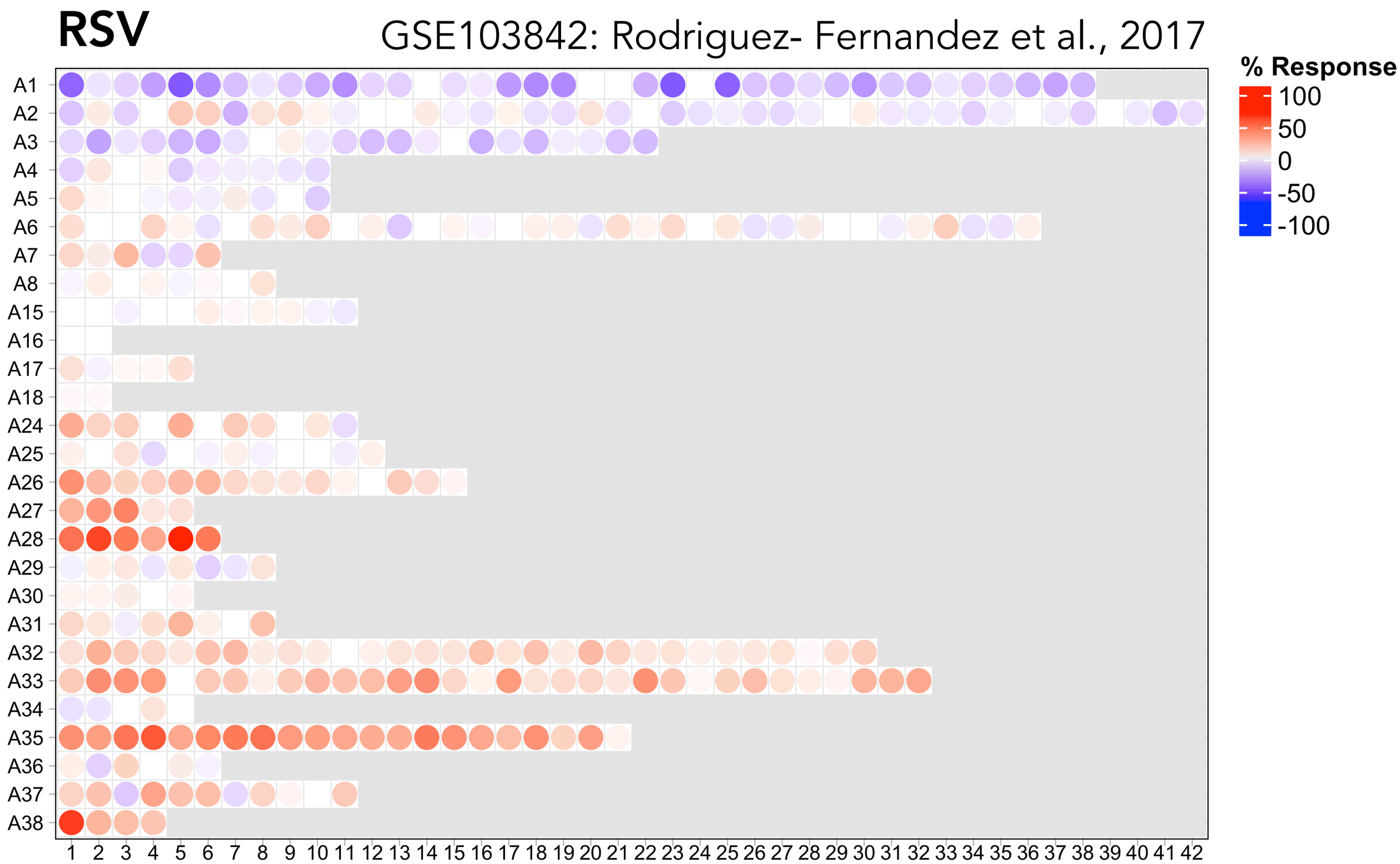
Figure 7: Individual RSV and control subjects PCA plots. Module-level data obtained for individual subjects from each of the six different studies were combined in a single PCA analysis. The overall number of subjects was 490, including 319 RSV and 171 controls. A 3D PCA plot is shown in the upper panel. 2D PCA plots are shown in the lower panel with subjects color coded according to grouping information and dataset membership.

Figure 8: Individual RSV and control subjects tSNE plots. Module-level data obtained for individual subjects from each of the six different studies were combined in a single tSNE analysis (overall number of subjects was 490, including 319 RSV and 171 control subjects). Subjects are color coded according to grouping information in the upper panel and dataset membership in the lower panel.

TABLES

Table 1: Description of the public RSV blood transcriptome datasets.

GSE ID	Reference	Title	N Subjects	Age demogra-phics	Sample type	Platform	Cluster # (Figure 4)
42026	Herberg et al. / 23901082 / [19]	Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes	22 RSV & 33 Controls	Pediatric	Whole blood (PaxGene)	Illumina HumanHT-12 V3.0	2
38900	Mejias et al. / 24265599 / [20]	Genome-wide analysis of whole blood transcriptional response to Respiratory Syncytial Virus, Influenza and Rhinovirus lower respiratory tract infection (LRTI) in children	135 RSV & 39 controls	Pediatric	Whole blood (Tempus)	Illumina HumanHT-12 V4.0	2
80179	McDonald et al. / 27822537 / [21]	A simple screening approach to prioritize genes for functional analysis identifies a role for IRF7 in the control of RSV disease.	27 RSV & 52 controls	Pediatric	Whole blood (PaxGene)	Illumina HumanHT-12 V4.0	2
73072	Liu et al. / 26801061 / [22]	Host gene expression signatures of H1N1, H3N2, HRV, RSV virus infection in adults	20 subjects 2 time points	Adult	Whole blood (PaxGene)	Affymetrix GeneChip Human Genome U133A 2.0 Array	1
77087	de Steenhuijsen Piters et al. / 27135599 / [23]	Nasopharyngeal microbiota, host transcriptome and disease severity in children with respiratory syncytial virus infection	81 RSV & 23 controls	Pediatric	Whole blood (Tempus)	Illumina HumanHT-12 V4.0	1
103842	Rodriguez-Fernandez et al. / 29045741 / [12]	Respiratory Syncytial Virus Genotypes and Disease Severity in Young Children Hospitalized with Bronchiolitis	62 RSV infection & 12 controls	Pediatric	Whole blood (Tempus)	Illumina HumanHT-12 V4.0	1


RSV
GSE103842: Rodriguez- Fernandez et al., 2017
% Response
100
50
0
-50
-100
A1 A2 A3 A4 A5 A6 A7 A8 A15 A16 A17 A18 A24 A25 A26 A27 A28 A29 A30 A31 A32 A33 A34 A35 A36 A37 A38
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42

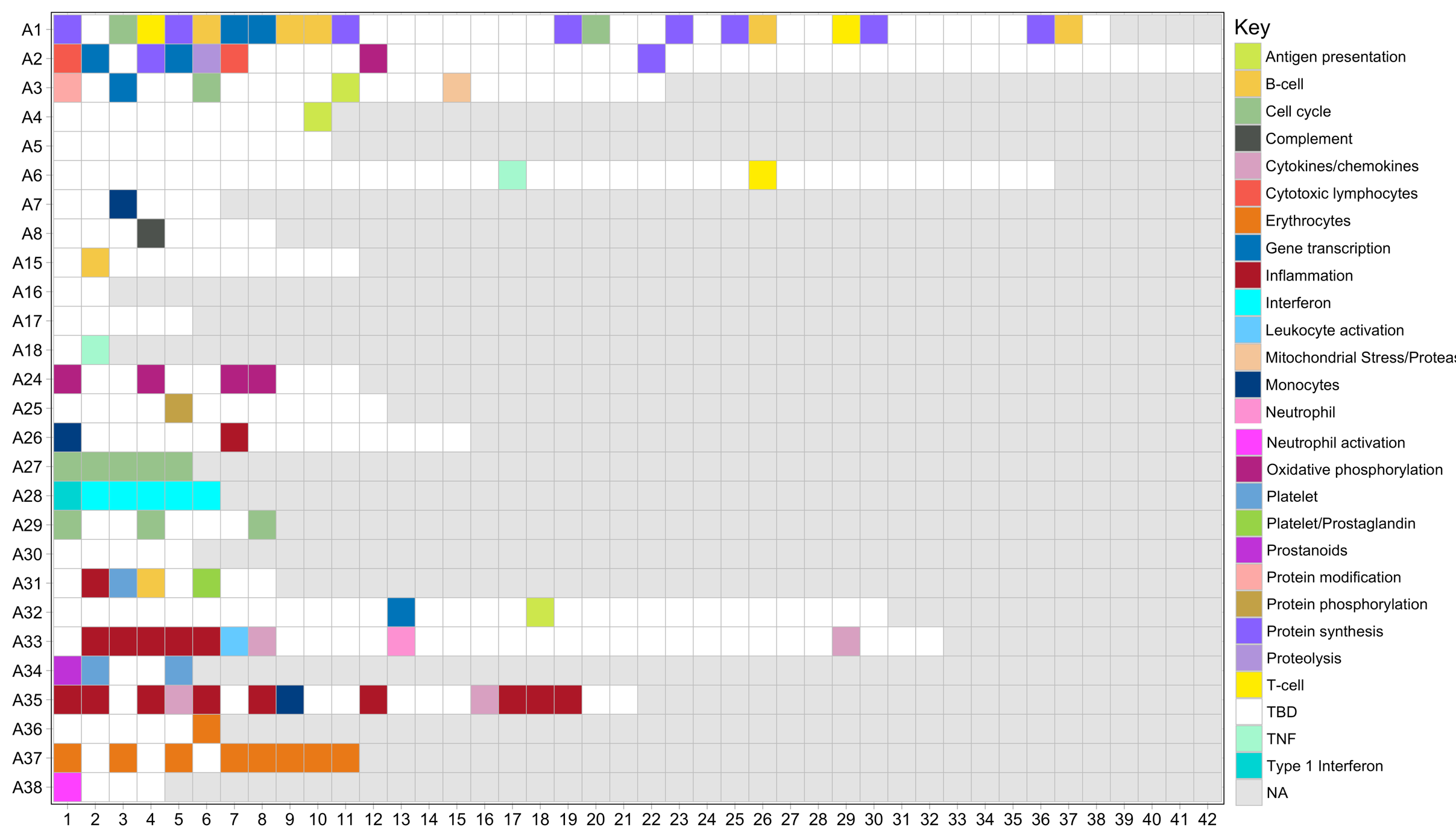

A1 A2 A3 A4 A5 A6 A7 A8 A15 A16 A17 A18 A24 A25 A26 A27 A28 A29 A30 A31 A32 A33 A34 A35 A36 A37 A38
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
Key
Antigen presentation
B-cell
Cell cycle
Complement
Cytokines/chemokines
Cytotoxic lymphocytes
Erythrocytes
Gene transcription
Inflammation
Interferon
Leukocyte activation
Mitochondrial Stress/Proteasome
Monocytes
Neutrophil
Neutrophil activation
Oxidative phosphorylation
Platelet
Platelet/Prostaglandin
Prostanoids
Protein modification
Protein phosphorylation
Protein synthesis
Proteolysis
T-cell
TBD
TNF
Type 1 Interferon
NA

Figure 1

RSV
GSE100150: Altman et al. 2019
% Response
100
50
0
-50
-100
Modules
Sub module
+
+
SLE
GSE100150: Altman et al. 2019
+
NC
Melanoma
GSE100150: Altman et al. 2019
NC
+
Influenza
GSE100150: Altman et al. 2019
+
NC
Transplant
GSE100150: Altman et al. 2019
+
+

Figure 2

RSV

OVER
UNDER
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% significant transcripts
cutoff p<0.05
GSE42026: Herberg et al.
GSE38900: Mejias et al.
GSE73072: Liu et al.
GSE77087: de Steenhuijsen Piters et al.
GSE80179: McDonald et al.
GSE103842: Rodriguez-Fernandez et al.

Figure 3

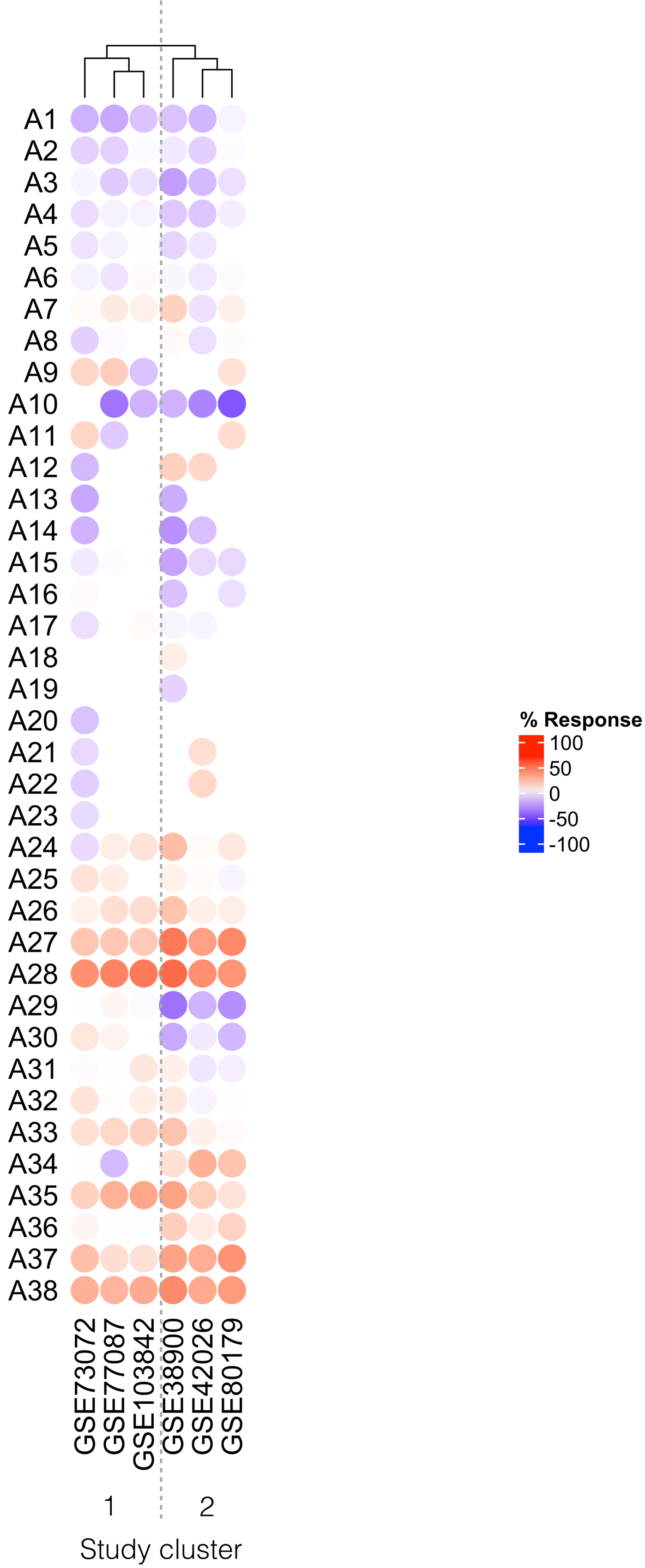

A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
A16
A17
A18
A19
A20
A21
A22
A23
A24
A25
A26
A27
A28
A29
A30
A31
A32
A33
A34
A35
A36
A37
A38
GSE73072
GSE77087
GSE103842
GSE38900
GSE42026
GSE80179
1
2
Study cluster
% Response
100
50
0
-50
-100

Figure 4

GSE103842: Rodriguez- Fernandez et al., 2017

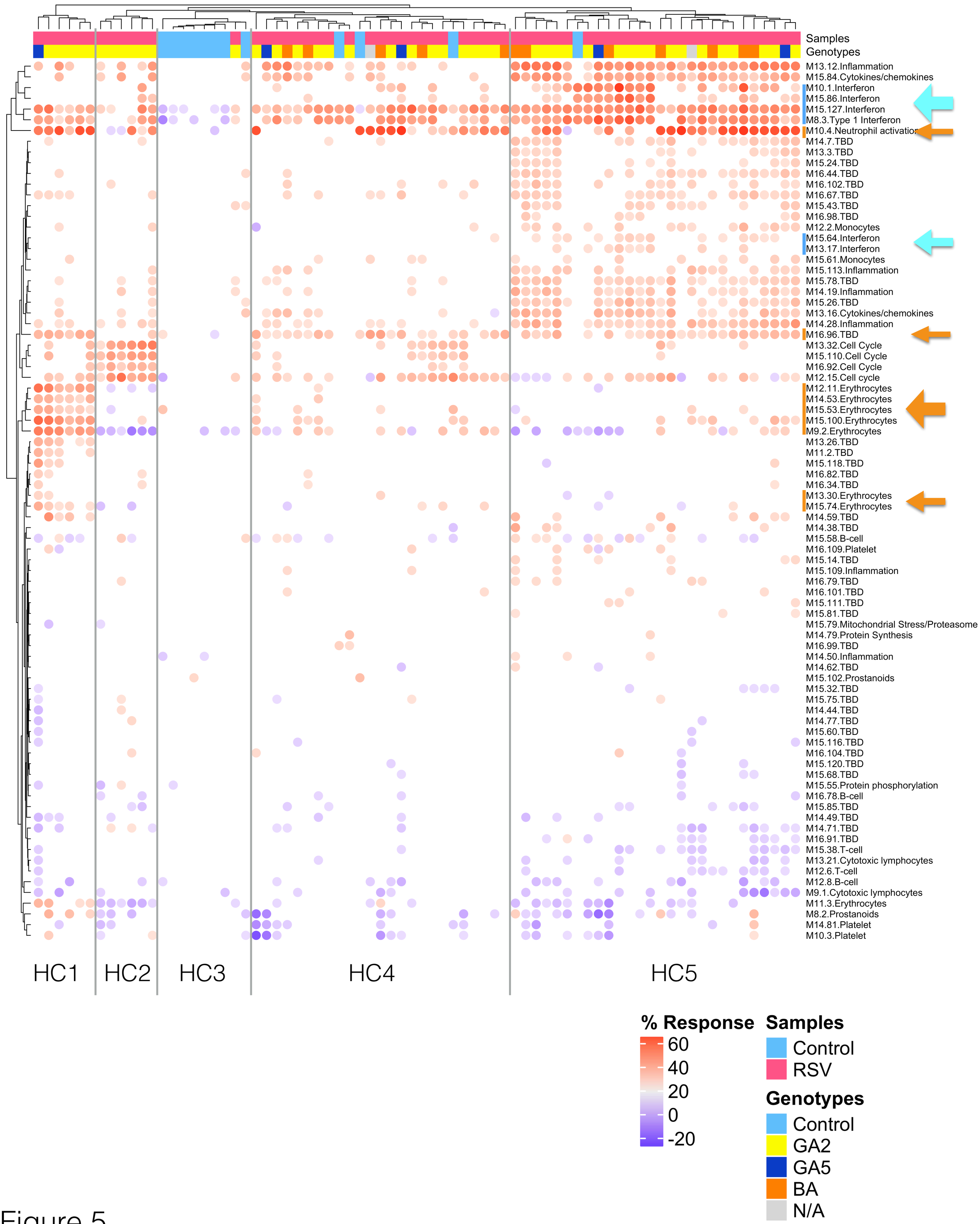
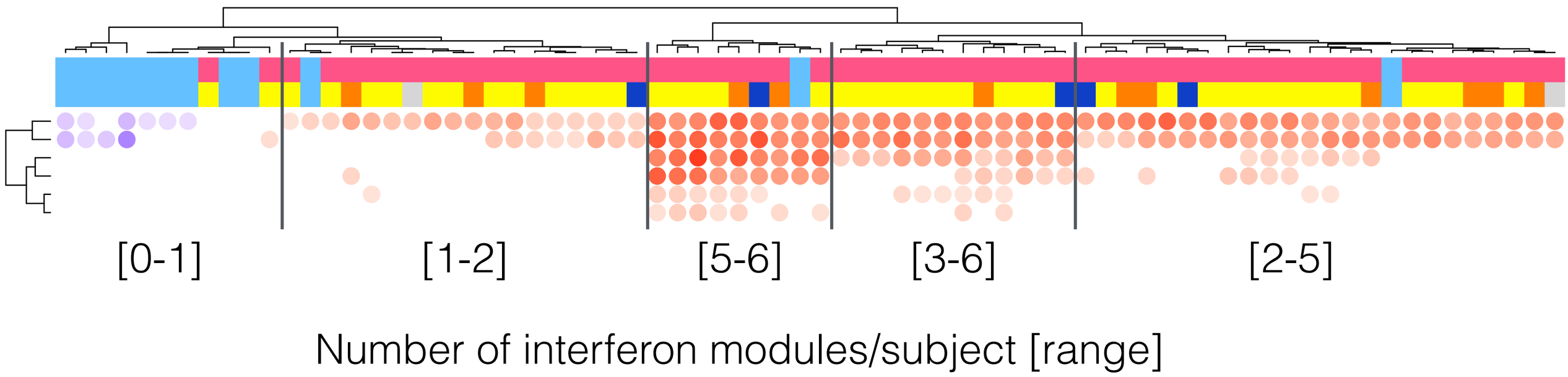
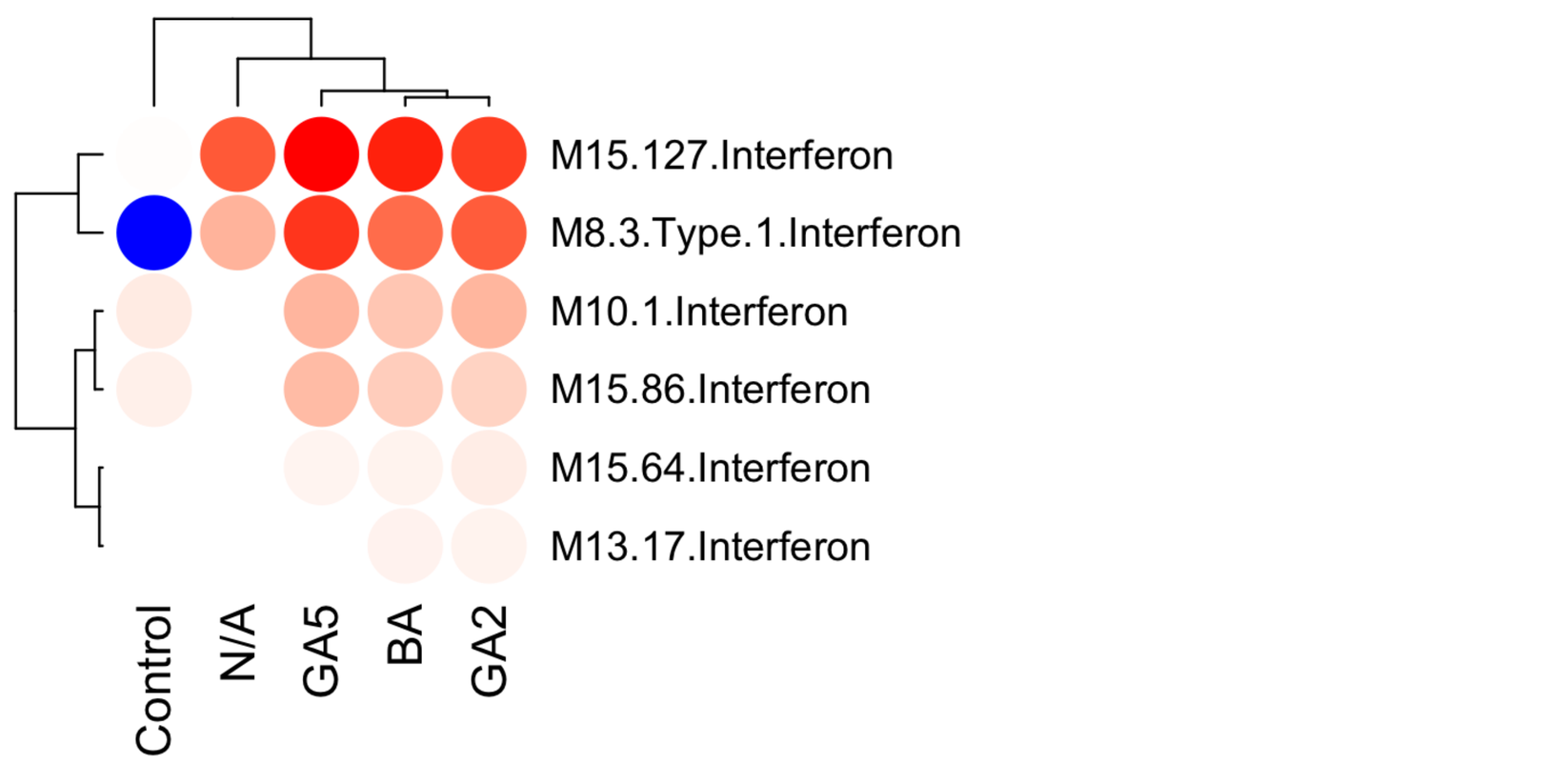


Figure 5

GSE103842: Rodriguez- Fernandez et al., 2017



Number of interferon modules/subject [range]

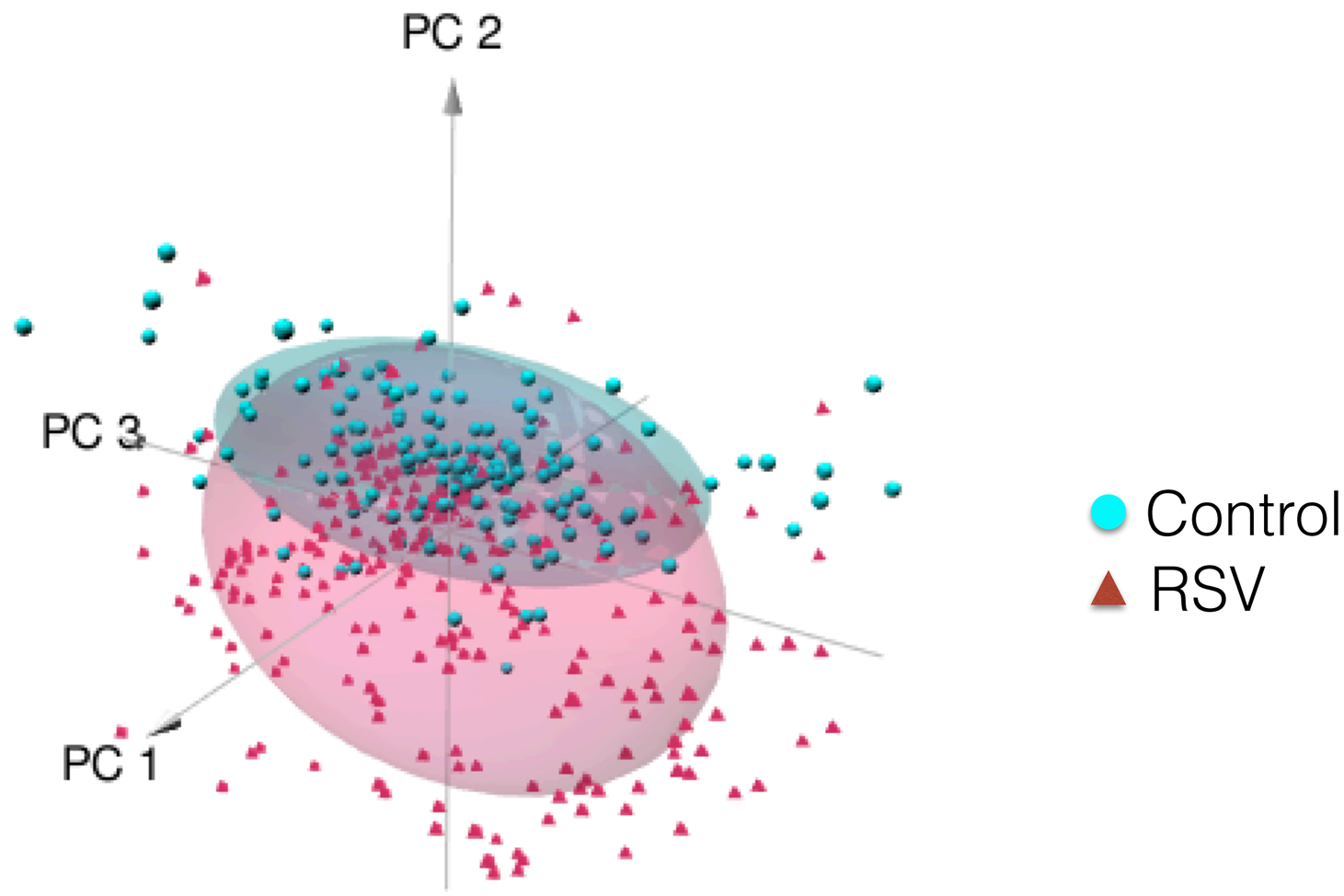


Interferon modules / RSV genotypes

Figure 6

PCA

3D Plot



2D Plots

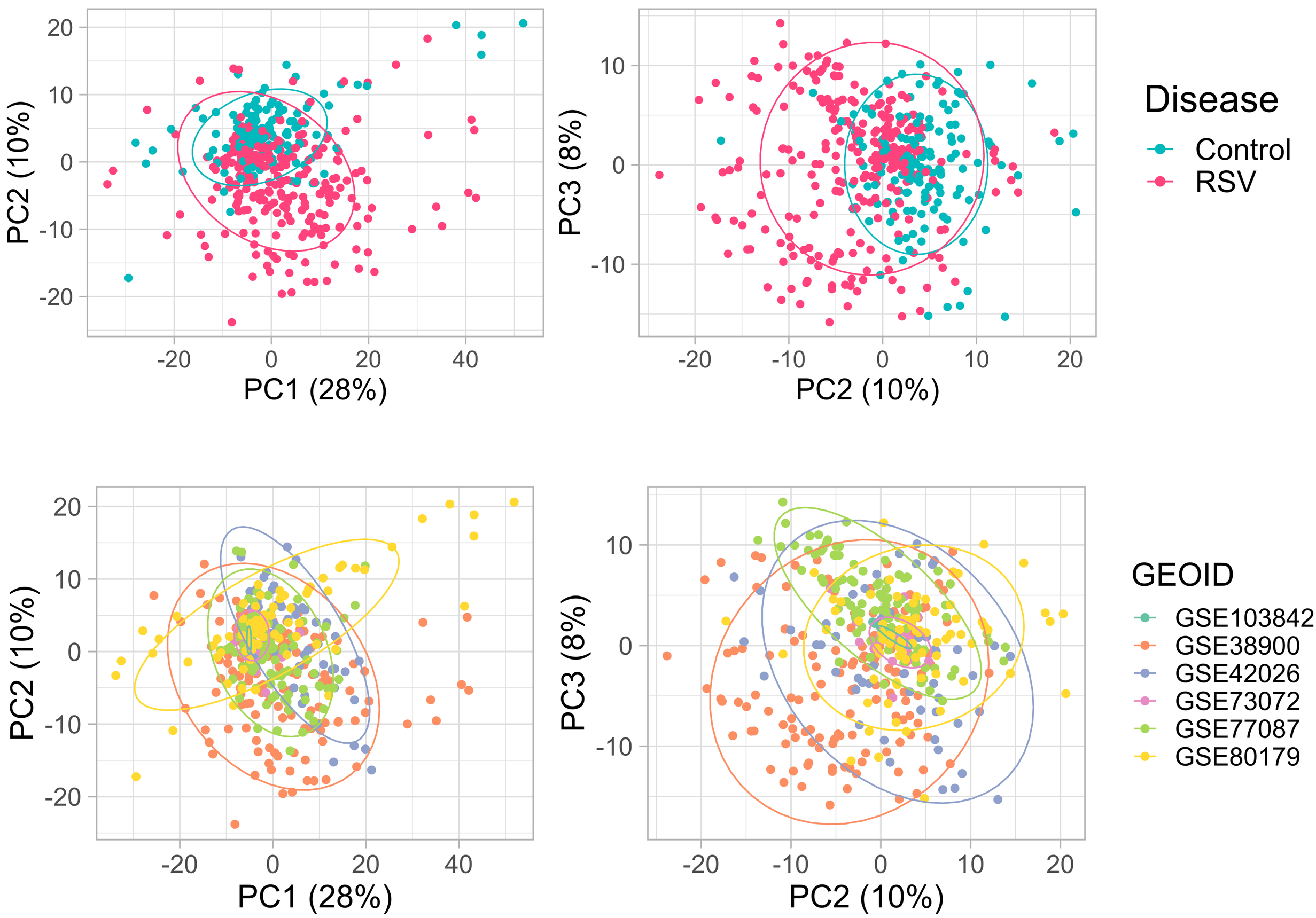
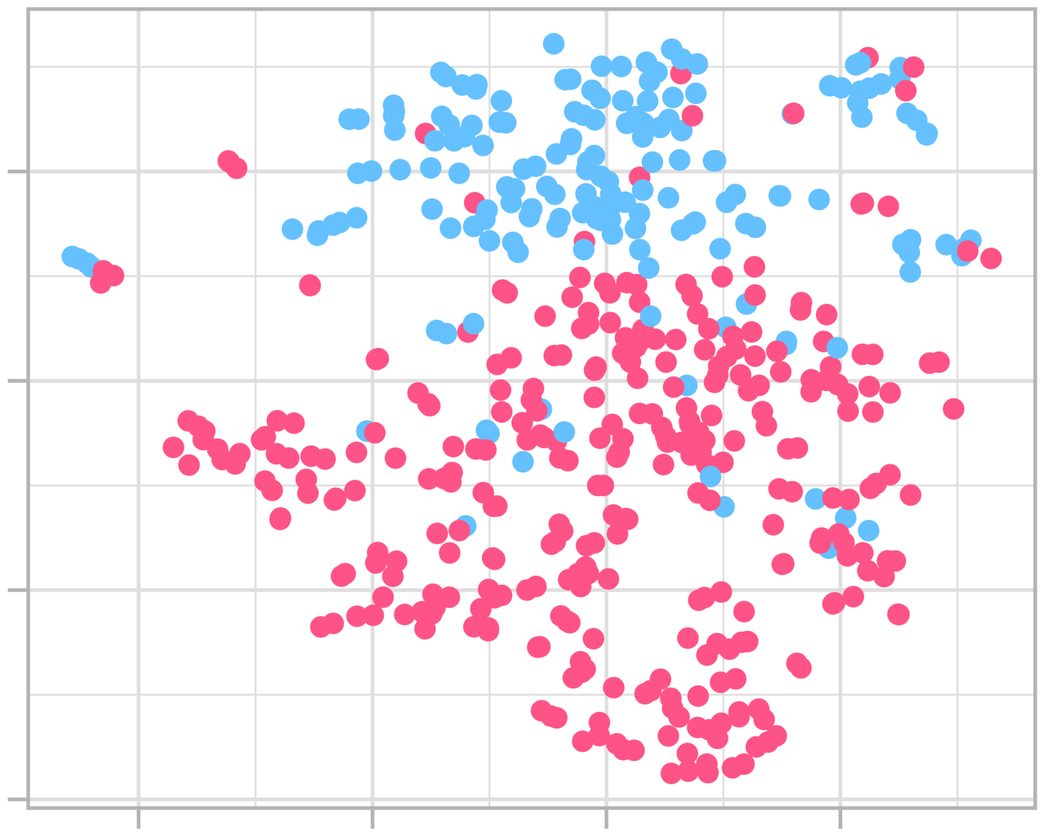


Figure 7

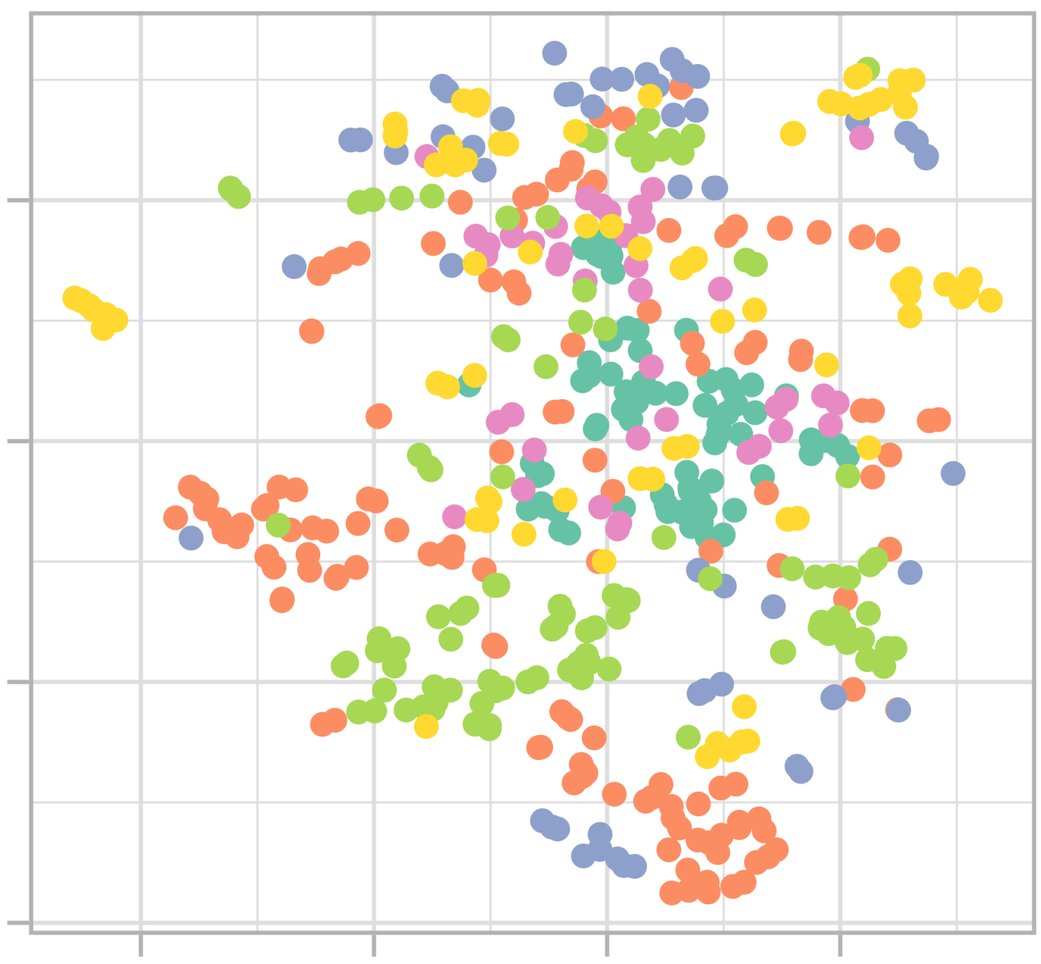
tSNE plots



by Group

Disease

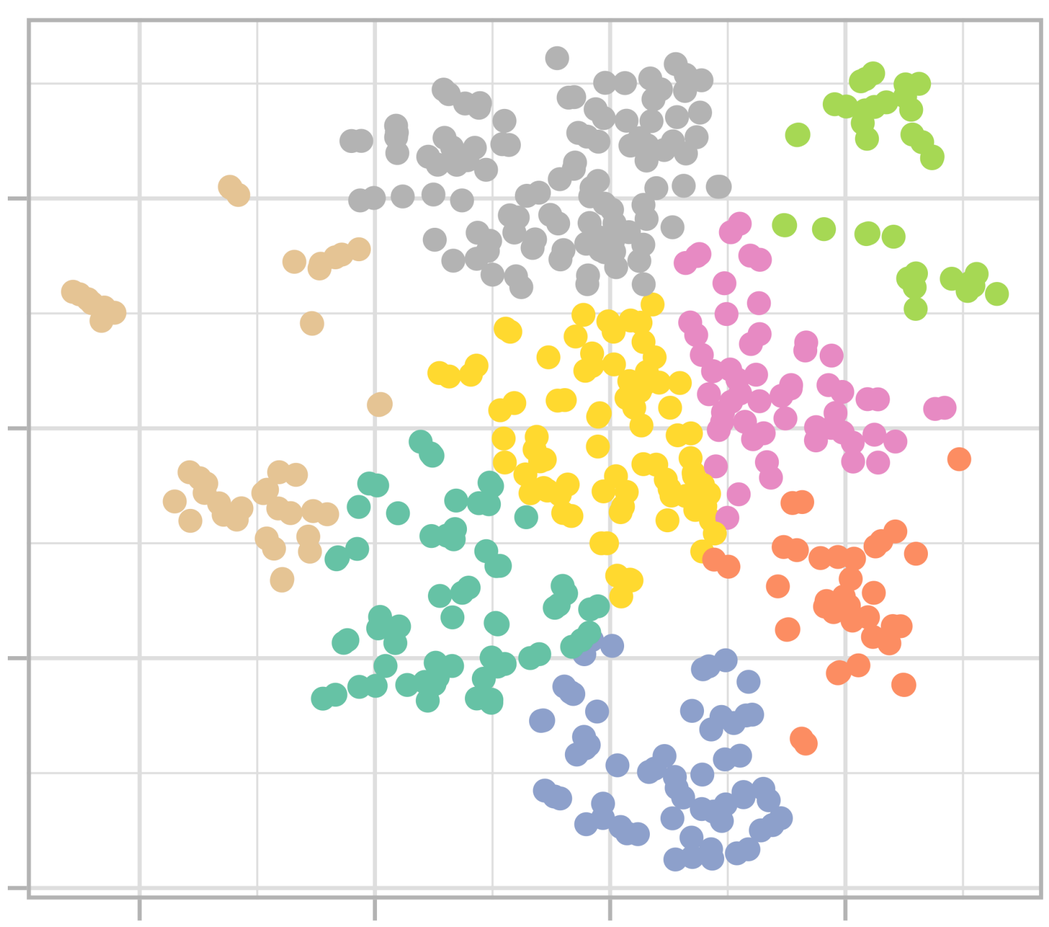
- Control
- RSV



by Dataset

GEOID

- GSE103842
- GSE38900
- GSE42026
- GSE73072
- GSE77087
- GSE80179



by Endotype

kmeans

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Figure 8